Appendix

Appendix A1.1 Study characteristics: Torgesen et al., 2006 (randomized controlled trial)

Characteristic	Description
Study citation	Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., et al. (2006). *National assessment of Title I interim report—Volume II: Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers*. Retrieved from Institute of Education Sciences, U.S. Department of Education Web site: http://www.ed.gov/rschstat/eval/disadv/title1interimreport/index.html
Participants	The study design was based on random assignment of 37 school units[1] to one of the four interventions, *Corrective Reading*, *Kaplan SpellRead*, *Failure Free Reading*, and *Wilson Reading System®*. Within each school, students were randomly assigned to the intervention condition or to the comparison condition.[2] This report focuses on eight school units assigned to *Wilson Reading System®*.[3] At the time of analysis, the study included a total of 71 third-grade students (53 in the intervention and 18 in the comparison groups). Sample size at posttest by outcome measure was not reported.[4] In the intervention group, 61% of the students were female, 45% were African-American, and 36% were eligible for the free/reduced lunch program. In the comparison group, 79% of the students were female, 32% were African-American, and 64% were eligible for the free/reduced lunch program.
Setting	Eight school units in Pennsylvania.
Intervention	*Wilson Reading System®* was implemented by nine teachers from November 2003 to May 2004. For purposes of this study only word-level skills were developed, although the complete version of *Wilson* contains instructional routines and materials that also focus on comprehension and vocabulary. A 50-minute lesson was delivered five days a week to groups of three students with various basic reading levels. The average capabilities of each three-student group determined the pace of learning. Many of the sessions took place during the students' regular classroom reading instruction but were held outside their regular classrooms. Thus intervention group students received less reading instruction in the classroom than did students in the comparison group. Implementation fidelity was examined by reading program trainers who observed the teachers and coached them over a period of months, project coordinators who observed a sample of instructional sessions, and ratings based on a sample of videotaped sessions. Implementation was rated as acceptable.
Comparison	The comparison group students received their regular reading instruction, which included typical classroom instruction and, in many cases, other services (such as another pull-out program). The comparison group students had fewer small group instructional hours than the intervention group students, but more one-on-one instructional hours.
Primary outcomes and measurement	The outcome measures in the alphabetics domain were the phonemic decoding efficiency and sight word efficiency subtests of the Test of Word Reading Efficiency (TOWRE) and the word identification and word attack subtests of the Woodcock Reading Mastery Tests–Revised (WRMT–R). The only measure in the fluency domain was the Oral Reading Fluency test. Measures in the comprehension domain were the passage comprehension subtest of the Group Reading Assessment and Diagnostic Evaluation (GRADE) and the passage comprehension subtest of WRMT–R. (See Appendix A2.1–2.3 for more detailed descriptions of outcome measures.)
Teacher training	Trainers from *Wilson Reading System®* provided teacher training, which included group instruction, coaching, telephone consultation, and independent study using the *Wilson Academy* online course. On average, intervention group teachers participated in 62.5 professional development hours across all phases of the study (initial training phase, practice phase, and implementation phase).

1. A school unit consists of several partnered schools so that the cluster included two third-grade and two fifth-grade instructional groups.
2. One of seven indicators of students' reading skills at baseline (TOWRE-SWE) showed statistically significant differences between the intervention and comparison groups. Baseline differences were taken into account in the WWC analysis of the program effects.
3. Findings on *Corrective Reading, Kaplan SpellRead,* and *Failure Free Reading* are included in other WWC beginning reading reports.
4. The study reported that four students in the intervention group and three students in the comparison group were lost to analysis. However, it is not clear whether those students were in third grade or were part of an additional sample of fifth-grade students also examined in this study. The fifth-grade sample included in this study is not reviewed in this report because it is outside the scope of the review. For sample relevancy criteria, please see the Beginning Reading Protocol.

Appendix A2.1 Outcome measures in the alphabetics domain

Outcome measure	Description
Phonics	
Test of Word Reading Efficiency (TOWRE): Phonetic Decoding Efficiency subtest	The TOWRE is a standardized, nationally normed measure. The phonetic decoding efficiency subtest measures the number of pronounceable printed nonwords that can be accurately decoded within 45 seconds (as cited in Torgesen et al., 2006).
TOWRE: Sight Word Efficiency subtest	The TOWRE is a standardized, nationally normed measure. The sight word efficiency subtest assesses the number of real printed words that can be accurately identified within 45 seconds (as cited in Torgesen et al., 2006).
Woodcock Reading Mastery Test–Revised (WRMT–R): Word Identification subtest	The word identification subtest is a test of decoding skills. The standardized test requires the child to read aloud isolated real words that range in frequency and difficulty (as cited in Torgesen et al., 2006).
WRMT–R: Word Attack subtest	This standardized test measures phonemic decoding skills by asking students to read pseudowords. Students are aware that the words are not real (as cited in Torgesen et al., 2006).

Appendix A2.2 Outcome measure in the fluency domain

Outcome measure	Description
Edformation Oral Fluency Assessment	This test measures the number of words correct per minute (WCPM) that students read using three brief grade-level passages (AIMSweb, as cited in Torgesen et al., 2006). These passages include both fiction and nonfiction text. The norms for this test are updated by Edformation each school year.

Appendix A2.3 Outcome measures in the comprehension domain

Outcome measure	Description
Reading comprehension	
Group Reading Assessment and Diagnostic Evaluation (GRADE): Passage Comprehension subtest	The GRADE is an untimed, norm-referenced standardized test. The passage comprehension subtest includes a passage of text and corresponding multiple-choice comprehension questions (as cited in Torgesen et al., 2006).
WRMT–R: Passage Comprehension subtest	In this standardized test, comprehension is measured by having students fill in missing words in a short paragraph (as cited in Torgesen et al., 2006).

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain[1]

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Wilson Reading System®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Wilson Reading System®* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Torgesen et al., 2006 (randomized controlled trial)[7]								
TOWRE: Phonetic Decoding Efficiency subtest	Grade 3	8/71	91.97 (15.00)	86.19 (15.00)	5.78	0.38	Statistically significant	+15
TOWRE: Sight Word Efficiency subtest	Grade 3	8/71	87.19 (15.00)	84.14 (15.00)	3.05	0.20	ns	+8
WRMT–R: Word Identification subtest	Grade 3	8/71	92.21 (15.00)	89.75 (15.00)	2.46	0.16	ns	+6
WRMT–R: Word Attack subtest	Grade 3	8/71	103.10 (15.00)	94.30 (15.00)	8.80	0.58	Statistically significant	+22
Domain average[8] for alphabetics						0.33	na	+13

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. The study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socio-economic status. The study found statistically significant positive effects on WRMT–R word attack scores at posttest only for students with initial high word attack scores and students with initial high PPVT scores. Finally, the study found statistically significant positive effects on WRMT–R word attack and TOWRE-PDE posttest scores only for students who were not eligible for free/reduced lunch program, but not for those students who were eligible for free/reduced lunch.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The intervention group mean is the comparison group mean plus the mean difference.
4. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
7. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Torgesen et al. (2006) and the alphabetics domain, no corrections for clustering were needed because students were assigned to conditions. Corrections for multiple comparisons were needed because the study's reported corrections for multiple comparisons were based on grouping of outcomes that differs from the grouping of domains for this review.
8. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A3.2 Summary of study findings included in the rating for the fluency domain[1]

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Wilson Reading System®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Wilson Reading System®* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Torgesen et al., 2006 (randomized controlled trial)[7]								
Oral Reading Fluency	Grade 3	8/71	46.95 (39.20)	41.00 (39.20)	5.95	0.15	ns	+6
Domain average[8] for fluency						0.15	ns	+6

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. The study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socio-economic status. No differences were found between subgroups of students for the outcome in the fluency domain.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The intervention group mean is the comparison group mean plus the mean difference.
4. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
7. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Torgesen et al. (2006) and fluency, no corrections for clustering were needed because students were assigned to conditions. No corrections for multiple comparisons were needed because there is only one outcome in this domain.
8. This row provides the domain average, which in this instance is also the single outcome finding from the one study.

Appendix A3.3 Summary of study findings included in the rating for the comprehension domain[1]

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Wilson Reading System®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Wilson Reading System®* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Torgesen et al., 2006 (randomized controlled trial)[7]								
GRADE: Passage Comprehension subtest	Grade 3	8/71	89.97 (15.00)	85.78 (15.00)	4.19	0.28	ns	+11
WRMT-R: Passage Comprehension subtest	Grade 3	8/71	93.87 (15.00)	92.87 (15.00)	1.00	0.07	ns	+3
Domain average[8] for comprehension						0.17	ns	+7

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. The study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socioeconomic status. No differences were found between subgroups of students for outcomes in the comprehension domain.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The intervention group mean is the comparison group mean plus the mean difference.
4. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
7. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Torgesen et al. (2006) and the comprehension domain, no corrections for clustering were needed. No correction for multiple comparisons were needed because the study's reported corrections for multiple comparisons were based on the same grouping of outcomes as the domain for this review.
8. This row provides the domain average, which in this instance is also the study average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A4.1 *Wilson Reading System®* rating for the alphabetics domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of alphabetics, the WWC rated *Wilson Reading System®* as potentially positive effects. It did not meet the criteria for positive effects because only one study showed a statistically significant positive effect. The remaining ratings (mixed, no discernible effects, potentially negative, or negative) were not considered because *Wilson Reading System®* was assigned the highest applicable rating.

Rating received

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Met. One study showed a statistically significant positive effect.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Met. No studies showed a statistically significant or substantively important negative effect. The single study that met the WWC standards showed a statistically significant positive effect.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. Only one study showed a statistically significant positive effect.

AND

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. No studies showed statistically significant or substantively important negative effects.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A4.2 *Wilson Reading System*® rating for the fluency domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of fluency, the WWC rated *Wilson Reading System*® as no discernible effects. It did not meet the criteria for other ratings (positive effects, potentially positive effects, mixed effects, potentially negative effects, and negative effects) because the single study that met WWC standards did not show statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either *positive* or *negative.*

  Met. No studies showed statistically significant or substantively important positive or negative effects.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed statistically significant positive effects.

AND

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. No studies showed statistically significant or substantively important negative effects.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed statistically significant or substantively important positive effects.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. The single study that met WWC standards showed indeterminate effects.

Mixed effects: Evidence of inconsistent effects as demonstrated through either of the following criteria.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed statistically significant or substantively important effects, either positive or negative.

OR

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Not met. No studies showed statistically significant or substantively important effects, either positive or negative.

(continued)

Appendix A4.2 *Wilson Reading System®* rating for the fluency domain *(continued)*

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

  Not met. No studies showed statistically significant or substantively important negative effects.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

  Met. No studies showed statistically significant or substantively important positive effects. In addition, no studies showed a statistically significant or substantively important negative effect.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed statistically significant negative effects.

AND

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

  Met. No studies showed statistically significant or substantively important positive effects.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A4.3 *Wilson Reading System*® rating for the comprehension domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of comprehension, the WWC rated *Wilson Reading System*® as no discernible effects. It did not meet the criteria for other ratings (positive effects, potentially positive effects, mixed effects, potentially negative effects, and negative effects) because the single study that met WWC standards did not show statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either *positive* or *negative.*

  Met. No studies showed statistically significant or substantively important positive or negative effects.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed statistically significant positive effects.

AND

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. No studies showed statistically significant or substantively important negative effects.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed statistically significant or substantively important positive effects.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. The single study that met WWC standards showed indeterminate effects.

Mixed effects: Evidence of inconsistent effects as demonstrated through either of the following criteria.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed statistically significant or substantively important effects, either positive or negative.

OR

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Not met. No studies showed statistically significant or substantively important effects, either positive or negative.

(continued)

Appendix A4.3 *Wilson Reading System®* rating for the comprehension domain *(continued)*

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

  Not met. No studies showed statistically significant or substantively important negative effects.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

  Met. No studies showed statistically significant or substantively important positive effects. In addition, no studies showed a statistically significant or substantively important negative effect.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed statistically significant negative effects.

AND

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

  Met. No studies showed statistically significant or substantively important positive effects.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		School units	Students	
Alphabetics	1	8	71	Small
Fluency	1	8	71	Small
Comprehension	1	8	71	Small
General reading achievement	0	0	0	na

na = not applicable/not studied

1. A rating of "moderate to large" requires at least two studies and two schools across studies in one domain, and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small."